

DOCUMENT RESUME

ED 172 946

PS 010 721

AUTHOR Colby, Anne; And Others
 TITLE A Longitudinal Study of Moral Judgment.
 PUB DATE Mar 79
 NOTE 17p.; Paper presented at the Biennial Meeting of the Society for Research in Child Development (San Francisco, California, March 15-18, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
 DESCRIPTORS *Developmental Stages; Longitudinal Studies; *Moral Development; Reliability; Research Methodology; Validity; *Young Adults; *Youth

ABSTRACT

Findings of this 20-year longitudinal study support Kohlberg's theoretical predictions of invariant structural consistency of subjects' responses to hypothetical moral dilemmas. Subjects were 58 American males chosen in 1955 according to age, social class and sociometric status. They were 10, 13, or 16 years of age at Time 1. Half of each age group were upper middle class, half working class. Half the subjects in each age and SES group were sociometric stars, half were sociometric isolates. Subjects were interviewed every 4 years between 1956 and 1977. Each interview included the presentation of nine hypothetical moral dilemmas with follow-up questions to probe the subject's reasoning. The Standard Form Scoring Manual (1978) was used to stage score the interviews. All scoring was done blind for each individual dilemma. A central point of the longitudinal analysis was to demonstrate the reliability and validity of the scoring system. If a stage is a structural whole, the individual should be consistent over various stimuli and occasions for testing. Reliability data are consistent with this requirement. Results indicate that in no case on either form did a subject reach a stage without having gone through each preceding stage. Almost all of the scores assigned an interview were at two adjacent stages. People seem in their development to go through certain qualitatively distant stages of moral reasoning. Development seems to be highly regular from stage to stage and highly consistent at each stage level. (Author/RH)

 * Reproductions supplied by EDRS are the best that can be made *
 * from the original document. *

A LONGITUDINAL STUDY OF MORAL JUDGMENT

Anne Colby, Lawrence Kohlberg, and John Gibbs

Paper presented at SRCD, San Francisco
March, 1979

4/10/79
Final draft

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

John Gibbs

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

We are now ready in 1979 to present the results of a longitudinal study of moral judgment that Lawrence Kohlberg began in 1955. It has taken this long partly because a twenty year time span was needed to cover development through all five stages, partly because the stage definitions and scoring procedures have been evolving over the years. The longitudinal study has played a central role in the process of this evolution.

According to cognitive-developmental theory, one would expect each of Kohlberg's longitudinal subjects to move through the stages in the proposed order, without any downward stage movement or stage skipping. In fact, according to the original analysis in 1969, most subjects did follow this invariant sequence in their development through the stages, but there was also a substantial number of disconfirming cases. At this point, Kohlberg could have concluded that his initial assumptions had been supported within the limits of measurement error and then gone on to use the original stage descriptions and scoring system to continue his research. Instead, he interpreted the discrepancies as pointing to a need to refine the descriptions and scoring methods in order to bring theory, data, and measurement into closer agreement.

Since the early 1970s, we have been using a subset of the longitudinal cases to redefine the central features of the moral judgment developmental sequence and to construct a better scoring system. The reformulation process has involved a mutual bootstrapping between the definitional/descriptive and methodological components of the theory, with the core assumptions of cognitive-developmental theory (invariant sequence, structured wholeness, and so on)

serving as criteria against which to evaluate the revisions. Our hope was that this process would give us increasingly precise and accurate pictures of moral judgment development. We expected that when the remaining longitudinal cases were reanalyzed using the revised stage criteria, the data would fit the core assumptions of cognitive-developmental theory much more closely than did the data in the original analysis. This expectation seems to have been justified, as is shown by the results of our current analysis.

This evolutionary, bootstrapping process has resulted in moral judgment stage definitions which are more structural and a scoring system which is more reliable and valid than earlier, more intuitive or clinical systems. We have used this system to rescore all the longitudinal interviews. In this paper we shall focus on the two aspects of our results that are most critical from the point of view of cognitive development theory - internal consistency of moral reasoning (the structured wholeness of a stage) and invariant developmental sequence.

The subjects were 58 American males. They were chosen in 1955 according to age, social class, and sociometric status. They were 10, 13, or 16 years old at Time 1. Half of each age group was upper middle class, half was working class. Half the subjects in each age and SES group were sociometric stars, half were sociometric isolates.

Subjects were interviewed every four years between 1956 and 1977. Each interview included the presentation of nine hypothetical moral dilemmas with follow-up questions to probe the subject's reasoning. In our analysis, we have grouped the dilemmas into three Interview Forms: Form A, B, and C. Each form includes three dilemmas. The analysis of Form C has not yet been completed so this paper will report the results of Forms A and B. The Standard Form Scoring Manual (1978) was used to stage score the interviews. All scoring was done blind, by individual dilemma.

Before going on to look at the results of the longitudinal study, let us consider the reliability of the scoring system. This is important because a central point of the longitudinal analysis has been to demonstrate the reliability and validity of that system. The test-retest figures are especially important because they give us the most relevant estimate of measurement error. Since we could not expect our measure to be error free, we needed some estimate of measurement error against which to evaluate deviations from perfect sequentiality. Therefore, the analysis of longitudinal sequence involved a comparison of the frequency of sequence deviations (downward stage movement) in the longitudinal data with the frequency of negative Time 1 - Time 2 changes in our test-retest data. (Of course, this comparison depends upon a psychometrically adequate level of test-retest reliability.) If sequence deviations exceed test-retest instability, we cannot consider our data to support the invariant sequence assumption.

Table 1 summarizes test-retest, alternate form, and interrater reliability. The reliability data are presented several ways so some discussion of the table is needed. Correlations between Time 1 and Time 2 in test retest reliability data are presented for Forms A and B separately. Both are in the high 90's.

Since the correlations could be very high without much absolute agreement between scores at Time 1 and Time 2, we have also presented percent agreement figures.

You can see that for almost all subjects the scores on Times 1 and 2 were within a third of a stage of each other. (One step: from 1 to 1(2) or 1(2) to 2(1) or 2(1) to 2, and so on.) If we consider global scores based on a nine point scale - the five stages and the four transition points between stages, we find between 70 and 80% complete agreement. We also

calculated percent agreement using a more differentiated system of global scores with a thirteen-point scale. This system includes two transition points between each stage, distinguishing, for example between an interview which is primarily Stage 2 with some 3 versus an interview which is primarily Stage 3 with some 2. The agreement levels for this system were in the sixties for the separate forms, 70% for the two forms combined.

Percent absolute agreement figures for interrater reliability on Form A ranged from 88% to 100% for agreement within a third of a stage, from 75% to 88% for complete agreement based on the nine-point scale and from 53% to 63% for complete agreement using the most finely differentiated scale. The correlation for Raters 1 and 2 on the Form A test-retest interviews was .98. Interrater figures for Form B are about the same as those for Form A.

The alternate form figures are based on the Form A and B longitudinal scores. The correlation is .84. Since Form A was scored by Rater 1 and Form B by Rater 2, the reliability figures confound form and rater differences. Alternate form reliability was also calculated for those ten test-retest subjects who received both Forms A and B. A single rater scored independently both forms of the twenty interviews. Percent agreement between Forms A and B for this sample was comparable to test-retest and inter-rater reliability: 100% of the interviews were given scores within one third stage of each other for the two forms, 75% received identical scores for Forms A and B using the nine-point scale, and 67% received identical scores for the two forms using the thirteen-point scale. The correlation between moral maturity scores for Forms A and B in this sample was .95.

As we argued in our 1974 proposal to NICHD, the appropriate validity concept for a developmental measure such as the Standard Form is construct validity, not prediction to an external criterion. For a measure of moral judgment stage, the two most critical empirical criteria of construct validity correspond to the two most central theoretical assumptions of the cognitive developmental model. They are invariant stage sequence and internal consistency (generality of stage usage across moral issues or dilemmas). As noted in the following description of our longitudinal study, the data now analyzed confirm both invariant sequence and internal consistency. According to Kohlberg (1978), "Construct validity means the fit of the data obtained by means of the test to primary components of its theoretical definition. The primary theoretical definition of structural moral development is that of an organization passing through invariant sequential stages." In other words, positive results of the longitudinal analysis support not only the theoretical assumptions, but also the validity of the measure. Negative results, of course, could be due either to a bad theory, a bad test, or both. From the point of view of stage theories like Piaget's or Kohlberg's, test reliability and test construct validity are one and the same thing. Cronback and Glaser (1972) have pointed out that the validity and reliability of a test are basically related notions, both refer to the generalizeability of performance on a test, or a set of test items, to performance in other situations including the performance on other forms of the test or other times of testing. In the case of structural stage, construct validity demands high test-retest and alternate-form reliability. If a stage is a structural whole, the individual should be consistent over various stimuli and occasions for testing. Our reliability data fit this demand rather well.

Turning to the longitudinal sample, let us first consider the sequence

data. In Form A sequence reversal or downward stage changes occurred in 7% of the adjacent time points (using our most differentiated 13 point scale of global interview scores). The reversals in Form B were 8% and in Forms A and B combined 7%. A comparison with downward stage change in test-retest data is presented in Table 2. You can see that in every case the test-retest reversals are well over twice as great as the longitudinal reversals so it seems reasonable to attribute the violations of longitudinal sequence to measurement error.

The concept of invariant stage sequence also implies that no stage will be omitted as development proceeds. In fact, we found that in no case on either form did a subject reach a stage in the sequence without having gone through each preceding stage. For the most part, changes across the four year intervals were less than a full stage. In Forms A and B, only 3% changed more than a full stage in four years. Scores combined across both forms show only half a percent changing more than a full stage.

According to cognitive-developmental theory, the logic of each stage forms "a structured whole". In line with this assumption, we expected to find a high degree of internal consistency in stage scores assigned. The data support this assumption as clearly as they do that of invariant sequence. In Form A only 3 of the 240 interviews showed three stages of reasoning which exceeded 20% of the total scores assigned. In Form B none of the 222 interviews showed three stages. The mean percentage of reasoning at the individual's modal stage was 68% for Form A, 72% for Form B and 67% for the two combined. Percentage of reasoning at the first two stages was 98% for Form A, 97% for Form B, and 95% for the two combined. In every case but one or two the two were adjacent stages. It is clear, then, that almost all of the scores assigned an interview were at two adjacent stages. (Remember that there are three dilemmas per form and each dilemma was scored without

knowledge of responses to the other dilemmas, so these figures cannot be an artifact of scorer bias.)

Analyses of age, SES, and sociometric status, should be completed soon. We also plan to use the longitudinal data to perform an item analysis of the Standard Form Scoring System, to investigate moral judgment substages, to ask some questions about the relation between content and structure in moral judgment, and so on. Those analyses will take somewhat longer. Before concluding we should point out one more thing about the longitudinal scores. In analyzing our data we found a certain amount of Stage 5 - 14% of our subjects reached at least the 4-5 transition - but no trace of Stage 6. In fact, we have dropped Stage 6 from the scoring system, partly because we found nothing in the longitudinal sample which seemed intuitively to be Stage 6, partly because the Standard dilemmas are not ideal for differentiating between Stages 5 and 6.

At least within the limits of this American male sample, then, our data strongly support the theoretical predictions of invariant stage sequence and internal structural consistency of subjects' responses to hypothetical moral dilemmas. People do seem in their development to go through certain qualitatively distinct stages of moral reasoning, and their development through these stages seems to be highly regular from stage to stage and highly consistent at each stage level.

Table 1

Reliability of Standard Form Scoring

Test-retest reliability

Correlation $T_1 - T_2$: Form A: .96 (Rater 1); .99 (Rater 2)
Form B: .97 (Rater 2)

Percent agreement within one-third stage: Form A 93%
Form B 94%
Forms A&B 100%

Percent agreement using pure stage and mixed stage scores
[9 categories: 1, 1/2, 2, 2/3, 3, 3/4, 4, 4/5, 5]

Form A 70% (Rater 1), 77% (Rater 2)
(N=43)

Form B 75% (Rater 2)
(N=31)

Forms A&B 80% (Rater 2)
(N=10)

Percent agreement using major/minor stage differentiations
[13 categories: 1, 1(2), 2(1), 2, 2(3), 3(2), 3, 3(4),
4(3), 4, 4(5), 5(4), 5]

Form A 59% (Rater 1), 70% (Rater 2)
Form B 62% (Rater 2)
Forms A&B 70% (Rater 2)

Alternate Form reliability (N=193)

Correlation Form A - Form B = .84
85% agreement within 1/2 stage (other percent agreement
figures not yet available)

Inter rater Reliability

Correlation - raters 1 and 2, Form A test-retest interviews
= .98

Percent agreement:

		within 1/3 stage	complete agreement (9 categ.)	complete agreement (13 categ.)
FORM A Rater pair	1	100	88	53
	2	100	88	63
	3	100	75	63
	4	88	88	63
	5	88	88	63
FORM B Rater pair	6	100	78	78

Table 2
Longitudinal Sequence

Subject and testing time		Global score Form A	Global score Form B	Global score Forms A&B combined
1-1		*1	2	2(1)
1-2		1(2)	2	2(1)
1-3		2(3)	*2	2(3)
1-4		3(2)	3	3
1-6		4(3)	3(4)	4(3)
Construction case	2-1	2(1)	2(1)	2(1)
	2-2	3	3	3
	2-3	4(3)	3(4)	3/4
	2-4	3 X	2(3) X	3(2) X
	2-5	5(4)	4(5)	5(4)
	2-6	5(4)	5(4)	5(4)
3-1		2(1)	2	2
3-2		2(3)	2(3)	2(3)
3-3		4(3)	3	3(4)
3-4		4(3)	3(4)	3(4)
3-5		4(3)	4(3)	4(3)
3-6		4(3)	4	4
4-1		2(3)	2	2
4-2		3	2(3)	3(2)
4-3		3(4)	3(2)	3
4-4		3(4)	*3(2)	3(4)
4-5		3(4)	4(3)	3(4)
5-1		2	1(2)	2(1)
5-2		2(3)	2	2(3)
5-3		2(3)	2(1) X	2 X
5-4		3 X	2	3(2)
5-5		3(2)	2	2(3) X
6-1		-	2(1)	2(1)
6-2		-	*2/3	2/3
6-3		-	3	3
6-4		-	3	3
6-5		-	*2/3	2/3 X
8-1		2	2(1)	2(1)
8-2		2	1 X	2(1)
8-4		3(4)	3	3

X = reversal

* = scores were all guesses

Subject and testing time		Global score Form A	Global score Form B	Global score Forms A&B combined
Construction case	9-1	2(1)	2(1)	2(1)
	9-2	2(3)	2(3)	2(3)
	9-3	3(2)	3	3
	9-4	3	3	3
	9-5	3(4)	3(4)	3(4)
	9-6	4	*4(3)	4
11-1		1(2)	2	2(1)
11-2		3(2)	*2	2(3)
12-1		2	1(2)	2(1)
12-2		3	2	3(2)
12-3		3(4)	3(4)	3(4)
12-4		3(4)	3	3(4)
13-1		2(1)	2	2
13-2		2	2	2
13-5		3	3	3
14-1		3(2)	2	2(3)
14-2		2(4) (3) X	3(2)	3(2)
14-4		4	4	4
14-5		-	4(5)	4(5)
15-1		2(3)	2(3)	2(3)
15-2		3(2)	3	3(2)
16-1		3(2)	2(3)	2(3)
16-2		3(4)	3	3
16-3		4(3)	4(3)	4(3)
16-4		4	4	4
16-6		4	4	4
Construction case	17-1	-	2	2
	17-2	3(2)	2/3	3(2)
	17-3	3(4)	3	3
	17-4	3(4)	3	3(4)
	17-5	4	4	4
	17-6	4(5)	5(4)	4(5)
18-1		3	3	3
18-2		3(2)	3	3
18-4		3/4	3(4)	3(4)
18-5		3/4	*3/4	3/4
18-6		4	4	4

Subject and testing time		Global score Form A	Global score Form B	Global score Forms A&B combined
19-1		2	2(1)	2(1)
19-2		2/3	3(4)	3
19-4		3	3(4)	3(4)
19-5		3(4)	4(3)	4(3)
19-6		4(3)	4	4
21-1		2	*2(1)	2
21-2		3(2)	2	2(3)
21-3		3	2/3	3(2)
21-4		3(2) X	3	3
22-1		2	2(1)	2
22-2		2(1) X	2(3)	2(3)
22-3		4(3)	3(4)	4(3)
22-4		4(3)	4	4
22-5		4(3)	-	4(3) X
22-6		4	*4(5)	4
Construction case	23-1	2	2(1)	2
	23-2	2/3	3(4)	3(4)
	23-3	3/4	4(3)	4(3)
	23-4	4(3)	4(5)	4
	23-5	4(5)	5	5(4)
	23-6	5(4)	-	5(4)
24-1		3	1(2)	2(1)
24-2		3	3(2)	3(2)
24-3		3(4)	3	3
24-4		3(4)	3(4)	3(4)
25-1		3(2)	2(1)	2(3)
25-2		3(4)	3	3
25-5		4(3)	4(3)	4(3)
25-6		4(3)	4	4
26-1		2	2	2
26-2		3	*3(2)	3
26-4		3	3	3
26-5		3	*3(4)	3
26-6		3(4)	3(4)	3(4)
27-1		2(3)	1(3)	3(2)
27-2		3(4)	3	3
27-4		4(3)	3	3(4)
27-5		3(4) X	3(4)	3(4)

Subject and testing time	Global score Form A	Global score Form B	Global score Forms A&B combined
29-1	2(3)	2(3)	2(3)
29-2	3	3	3
29-4	4(3)	3	3(4)
29-5	4	4(3)	4(3)
31-1	1(2)(3)	2	2(1)
31-2	3	2(3)	3(2)
31-5	3(4)	3(4)	3(4)
32-1	2(1)	2	2
32-2	2(3)	2	2
32-5	3	4(3)	3(4)
32-6	4(3)	4(3)	4(3)
36-1	2(3)	2(1)	2
36-2	2(1)(3) X	2	2
36-5	3(4)	3	3(4)
36-6	4(3)	3(4)	4(3)
37-1	3	3(2)	3
37-2	3(4)	3(4)	3(4)
37-3	*4(3)	3(4)	3(4)
37-4	4	*4(5)	4
37-5	4(5)	5(4)	5(4)
37-6	-	5(4)	5(4)
38-1	2(3)	2	2
38-2	3(4)	3(4)	3(4)
38-5	3/4	4	4(3)
38-6	3/4	3(4) X	3(4) X
39-1	2(3)	2(3)	2(3)
39-2	3(4)	3	3(4)
39-3	3(4)	3	3 X
39-4	4	4(3)	4
41-1	2/3	2	2
41-2	3(2)	3(2)	3(2)
41-3	3	3	3
41-4	3	3	3
41-5	3	3	3
41-6	4(3)	-	4(3)
Construction case 42-1	3(2)	3(2)	3(2)
42-2	3	3	3
42-4	4(5)	4	4(5)
42-5	5(4)	4(5)	4(5)
42-6	5(4)	5(4)	5(4)

Subject and testing time	Global score Form A	Global score Form B	Global score Forms A&B combined
43-1	*3	3(2)(1)	3
43-2	3(4)	3/4	3(4)
43-3	4(3)	3/4	4(3)
43-4	4(3)	4(3)	4(3)
44-2	2(4)	3	3
44-3	4(5)	4(3)	4
44-4	4 X	4	4
44-6	4(5)	4(5)	4(5)
45-1	3(2)	2	2(3)
45-2	3	3(2)	3
45-4	4	4(3)	4(3)
45-5	4	4	4
45-6	4	4	4
47-1	2(3)	3(4)	3(2)
47-2	3(2)	3/4	3(4)
47-3	3	-	3 X
47-4	3	3(4)	3(4)
47-5	3(4)	3(4)	3(4)
47-6	4(3)	2(4) X	4
48-1	2(3)	-	2(3)
48-2	3(4)	3(2)	3
48-4	4(3)	3	3(4)
48-5	4(3)	4(3)	4(3)
48-6	4(3)	4	4
49-1	3(4)	3	3
49-4	4(3)	4(3)	4(3)
49-5	3(4) X	3 X	3(4) X
49-6	3(4)	4(3)	3(4)
50-1	3	3(2)	3
50-4	3(4)	3	3
50-5	4(3)	4	4
50-6	4(3)	4	4
51-1	2	2	2
51-4	3(4)	3(4)	3(4)
51-5	3/4	4	4(3)
51-6	3(4)	3(4) X	3(4) X
53-1	3	*3(1)(2)	3
53-2	3(4)	*3(2)(4)	3(4)
53-4	3 X	3(4)	3(4)

Subject and testing time	Global score Form A	Global score Form B	Global score Forms A&B combined
54-1	2(3)	2(3)	2(3)
54-5	3	3/4	3(4)
56-1	2	-	2
56-4	3	-	3
56-6	4(3)	-	4(3)
59-1	2(3)	1(2)	2(1)(3)
59-2	3	2/3	3(2)
59-4	3	3	3
59-5	3(4)	3(2) X	3
62-1	2(3)	3	3(2)
62-2	3(4)	4	4(3)
62-4	3 X	*3 X	3 X
62-6	4(3)	4	4
64-1	3	3(2)	3
64-4	3	3(4)	3(4)
64-5	3(4)	4(3)	4(3)
Construction case 65-1	3	3	3
65-2	4	*2 X	4
65-4	4(3) X	4(3)	4(3) X
65-5	4(3)	4(3)	4(3)
65-6	4(3)	4	4
Construction case 67-1	3	2(3)	3(2)
67-2	3	3	3
67-4	4(3)	3(4)	4(3)
67-5	4(5)	3(4)	4
67-6	5(4)	4(3)	4(5)
68-1	3	3	3
68-2	3(4)	3	3(4)
68-4	4(3)	3(4)	3(4)
68-5	4	*3(4)	4(3)
68-6	4	3 X	4(3)
70-1	2(3)	3(2)	3(2)
70-2	3(4)	3(4)	3(4)
70-4	3(4)	4(3)	4(3)
70-5	4	3(4) X	4(3)
70-6	3(4) X	4	4(3)

Subject and testing time	Global score Form A	Global score Form B	Global score Forms A&B combined
71-1	3	3	3
71-2	3(4)	3	3
71-4	4(3)	3(4)	3(4)
71-5	4(3)	3 X	3(4)
81-1	3	2(1)	3(2)
81-2	3	3(2)	3(2)
81-5	4(3)	3(4)	3(4)
91-3	3	4(3)	3(4)
91-4	3(2) X	3 X	3 X
91-5	3(4)	3	3
91-6	4	4	4
92-3	3(2)	3	3(2)
92-4	3(4)	3(4)	3(4)
92-5	3 X	*3(4)	3(4)
93-3	4(3)	3	3(4)
93-4	3(4)	3	3(4)
93-5	3(4)	3(4)	3(4)
93-6	4(3)	4	4
95-3	3	3(2)	3
95-4	4(3)	-	4(3)
95-5	3(4)	3	3 X
95-6	-	3	3
96-3	4(3)	3	3(4)
96-4	3 X	3	3 X
96-5	4(3)	3(4)	3(4)
96-6	3(4) X	3 X	3 X

Table 3

Comparison of downward stage change
longitudinal* and test-retest

Pure and mixed stage scores

	Form A (Rater 1)	Form B (Rater 2)	Forms A&B combined
Longit. $T_n \rightarrow T_{n+1}$	2% (3/177)	4% (6/160)	1% (2/171)
Test-retest $T_1 \rightarrow T_2$	19%	23%	10%

Major/minor and pure stage scores

	Form A	Form B	Forms A&B
Longit. $T_n \rightarrow T_{n+1}$	7%	8%	7%
Test-retest $T_1 \rightarrow T_2$	19%	33%	20%

* Only the longitudinal interviews which were scored blind are included in this analysis